

# Minimum Relative Entropy Distributions With a Large Mean Are Gaussian

Matteo Smerlak\*

*Perimeter Institute for Theoretical Physics, 31 Caroline St. N., Waterloo ON N2L 2Y5, Canada*

(Dated: May 27, 2016)

We consider the following frustrated optimization problem: given a prior probability distribution  $q$ , find the distribution  $p$  minimizing the relative entropy with respect to  $q$  such that  $\text{mean}(p)$  is fixed and large. We show that solutions to this problem are asymptotically Gaussian. As an application we derive an  $H$ -type theorem for evolutionary dynamics: the entropy of the (standardized) distribution of fitness of a population evolving under natural selection is eventually increasing.

Keywords: relative entropy, constrained optimization, limit theorem, Gaussian distribution, natural selection

## INTRODUCTION

Relative entropy (aka Kullback-Leibler divergence) is the central concept of information theory [1]. Given two probability distributions  $p$  and  $q$ , the relative entropy of  $p$  with respect to  $q$ ,

$$D(p||q) \equiv \int p(x) \ln \frac{p(x)}{q(x)} dx, \quad (1)$$

measures the difference in information content between the (prior) distribution  $q$  and the (posterior) distribution  $p$ . As a consequence of Jensen's inequality,  $D(p||q) \geq 0$  with equality iff  $p = q$ . When  $q$  is uniform and  $x$  is discrete (resp. continuous),  $D(p||q)$  reduces to (minus) the Shannon (resp. Gibbs) entropy  $S(p)$ .

As first articulated by Jaynes [2], minimizing  $D(p||q)$  with respect to  $p$  under constraints is a powerful epistemological principle, leading to robust predictions with minimal input. This inference rule can also be motivated purely axiomatically [3]. On top of its foundational position in statistical mechanics, the Jaynes minimum relative entropy principle has been successfully applied to countless practical problems in virtually all fields of science [4]. Relative entropy literally attracts human attention [5].

Here we consider the following version of Jaynes' problem: given a distribution  $q$  supported on the real line, find the distribution  $p$  such that  $D(p||q)$  is minimum under the constraint that

$$\text{mean}(p) = \mu \quad (2)$$

for some constant  $\mu$ . We show that, if the solution exists for any  $\mu$ , then this solution is asymptotically Gaussian as  $\mu \rightarrow \infty$ . Moreover the rate of convergence to the Gaussian is determined by the tail behavior of  $q$  in a simple, explicit way.

Our original motivation for investigating this problem is from evolutionary theory [6]. In this context one is interested in characterizing the evolution of a population's distribution of fitness as a function of time (or generation number). As we shall discuss in the second part of this paper, the asymptotic Gaussianity of mean-constrained

minimum relative entropy distributions implies an  $H$ -type theorem for evolution: provided the population is sufficiently large and diverse, the entropy of (standardized) fitness distributions is eventually increasing under natural selection. Another, more elementary application to driven Brownian motion is also given for illustrative purposes.

## MAIN RESULT

Given a probability distribution  $q$  over the real line, it is well known that the minimizer of  $D(p||q)$  under the constraint that the expected value of some function  $g(x)$  be fixed to some value  $\gamma$  is  $p_\gamma(x) = e^{\lambda_\gamma g(x)} q(x) / Z_\gamma$ , where the Lagrange multiplier  $\lambda_\gamma$  is determined self-consistently as a function of  $\gamma$  (and  $q$ ) and  $Z_\gamma$  is a normalizing factor. In particular, taking  $g(x) = x$  (*i.e.* fixing the *mean* of  $p$ ) gives the exponentially tilted distribution<sup>1</sup>

$$p_\mu(x) = e^{\lambda_\mu x} q(x) / \chi_q(\lambda_\mu). \quad (3)$$

Here  $\chi_q(\lambda)$  is the cumulant-generating function of the prior  $q$  and  $\mu$  is the fixed value of the mean of  $p_\mu$ . The multiplier  $\lambda_\mu$  is obtained as the implicit solution of

$$\mu = \chi'_q(\lambda_\mu) / \chi_q(\lambda_\mu) = \psi'_q(\lambda_\mu) \quad (4)$$

with  $\psi_q(\lambda) \equiv \ln \chi_q(\lambda)$  the cumulant-generating function of  $q$ . Clearly, the relations above make sense for any  $\mu$  only if  $q$  decays faster than exponential for  $x \rightarrow \infty$ . To parametrize this decay rate we assume that

$$-\ln \int_x^\infty q(x) dx \underset{x \rightarrow \infty}{\sim} Cx^\alpha \quad (5)$$

for some  $C > 0$  and  $\alpha > 1$ .<sup>2</sup> Under this condition, the Kasahara Tauberian theorem [8] states that

$$\psi_q(\lambda) \underset{\lambda \rightarrow \infty}{\sim} D\lambda^{\bar{\alpha}} \quad (6)$$

<sup>1</sup> Expression (3) is known alternatively as the canonical ensemble (statistical physics), Cramér transform (probability theory), natural exponential family (statistics), Esscher transform (actuarial science) of  $q$ .

<sup>2</sup> A weaker condition requires that the LHS of (5) be regularly varying at infinity with index  $\alpha > 1$  [7].

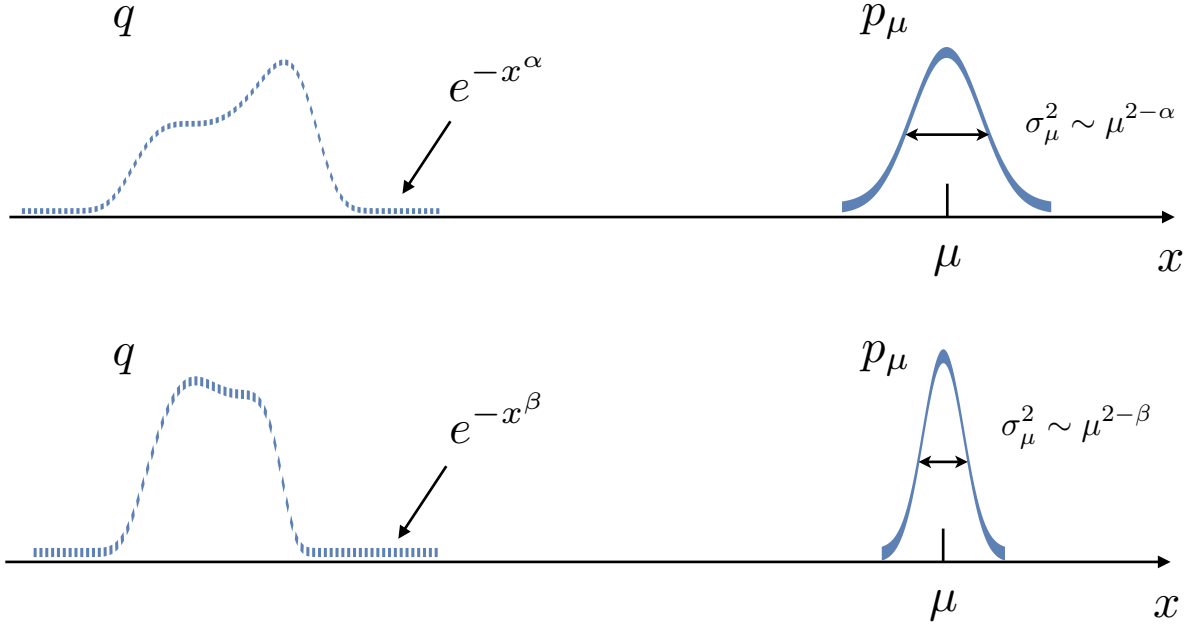


FIG. 1. Universality of mean-frustrated minimum relative entropy distributions. The minimizers  $p_\mu$  (continuous lines) of  $D(\cdot \| q)$  with a given (large) mean  $\mu$ , for two different priors  $q$  (dashed lines). All such minimizers are approximately Gaussian; the only feature distinguishing them is their variance  $\sigma_\mu^2$ , which is determined by  $\mu$  and the tail decay of  $q$  according to (11).

where  $\bar{\alpha} = \alpha/(1-\alpha)$  is the exponent conjugate to  $\alpha$  and  $D \equiv (\alpha C)^{-1/(\alpha-1)}/\bar{\alpha}$ . It follows that, in the limit where the mean  $\mu$  is large, we have

$$\lambda_\mu \underset{\mu \rightarrow \infty}{\sim} \alpha C \mu^{\alpha-1}. \quad (7)$$

Let us now show that in this limit  $p_\mu$  must be asymptotically Gaussian. Denote  $\sigma_\mu$  the standard deviation of  $p_\mu$  and let  $g_\mu(x) = \sigma_\mu p_\mu(\sigma_\mu x + \mu)$  be the standardized (*viz.* zero mean, unit variance) distribution associated to  $p_\mu$ . From (3), the  $j$ -th cumulant of  $g_\mu$  is given by

$$\kappa_\mu^{(j)} = \frac{\psi_q^{(j)}(\lambda_\mu)}{\psi_q''(\lambda_\mu)^{j/2}}. \quad (8)$$

Using the Kasahara theorem as above, we have

$$\psi^{(j)}(\lambda) \underset{\lambda \rightarrow \infty}{\sim} D(\bar{\alpha})_j \lambda^{\bar{\alpha}-j} \quad (9)$$

where  $(x)_j = x(x-1)\dots(x-j+1)$  denotes the falling factorial. It follows from (7) and (9) that the standardized cumulants  $\kappa_\mu^{(j)}$  with  $j \geq 3$  decrease increasingly fast as  $\mu \rightarrow \infty$ :

$$\kappa_\mu^{(j)} \underset{\mu \rightarrow \infty}{\sim} \frac{[(\alpha-1)C]^{1-j/2}(\bar{\alpha})_j}{(\bar{\alpha})_2^{j/2}} \mu^{\alpha(1-j/2)}. \quad (10)$$

In particular  $\kappa_\mu^{(j)} \rightarrow 0$  as  $\mu \rightarrow \infty$  whenever  $j \geq 3$ , *i.e.*  $g_\mu$  converges to the standard Gaussian distribution as announced. Moreover the variance of  $p_\mu$  is completely

determined by the tail behavior of  $q$  (and  $\mu$ ), as

$$\sigma_\mu^2 \underset{\mu \rightarrow \infty}{\sim} \frac{\mu^{2-\alpha}}{\alpha(\alpha-1)C}. \quad (11)$$

A uniform estimate of the rate of convergence can be obtained in terms of the relative entropy  $D(g_\mu \| \phi)$ ,<sup>3</sup> with  $\phi(x) \equiv (2\pi)^{-1/2} e^{-x^2/2}$ . Denoting  $\epsilon_\mu \equiv g_\mu - \phi$  we have

$$D(g_\mu \| \phi) \underset{\mu \rightarrow \infty}{\sim} \int \frac{\epsilon_\mu(x)^2}{\phi(x)} dx. \quad (12)$$

Now, we can write  $\epsilon_\mu$  from the cumulants  $\kappa_\mu^{(p)}$  by means of an inverse Laplace transform, yielding

$$\epsilon_\mu(x) \underset{\mu \rightarrow \infty}{\sim} \frac{(\bar{\alpha})_3 \phi(x)(x^3 - 3x)}{6[(\alpha-1)C]^{1/2}(\bar{\alpha})_2^{3/2}} \mu^{-\alpha/2}. \quad (13)$$

(A more general Edgeworth-type expansion [9] of  $\epsilon_\mu$  on the basis of Hermite polynomial follows similarly.) Plugging (13) into (12) gives

$$D(g_\mu \| \phi) \underset{\mu \rightarrow \infty}{\sim} \frac{(2-\alpha)^2}{6C\alpha(\alpha-1)} \mu^{-\alpha}. \quad (14)$$

Thus we see that, the thinner the tail of the prior distribution  $q$ , the faster the constrained minimizer  $p_\mu$  converges to the Gaussian attractor.

<sup>3</sup> Bounds on relative entropy are strong: by the Pinsker inequality, the total variation distribution  $\delta(p, q)$  between two distribution  $p$  and  $q$  is bounded as  $\delta(p, q) \leq \sqrt{D(p \| q)/2}$ .

We close this section by noting that (5) is certainly not the most general condition for  $p_\mu$  to be asymptotically Gaussian in the large mean limit. Consider for instance the thin-tailed Gumbel prior  $q(x) = \exp(-x - e^{-x})$ , a natural distribution in extreme value statistics [10]. Then we have  $\psi_q(\lambda) = \ln \Gamma(\lambda) \sim_{\lambda \rightarrow \infty} \lambda \ln \lambda$ , and repeating the computations above shows that  $D(g_\mu \| \phi) \rightarrow 0$  *exponentially* with  $\mu$ . (This example can be thought of as arising in the limit  $\alpha \rightarrow \infty$  of the above discussion.)

## REPRESENTATION AS TRANSPORT

It is interesting to consider the evolution of the shape of the minimizing distribution  $p_\mu$  when its constrained mean  $\mu$  is varied, or equivalently as the Lagrange multiplier  $\lambda$  is varied, as a dynamical system. It is straightforward to check that the minimizing solution  $p_\lambda(x) = e^{\lambda x} q(x) / \chi_q(\lambda)$  satisfies the integro-differential equation

$$\partial_\lambda p_\lambda(x) = (x - \mu_\lambda) p_\lambda(x). \quad (15)$$

Note that, in this dynamical perspective, the prior distribution  $q$  is just the initial condition  $p_0$  of the flow. Eq. (15) can be then used to derive an equation for the standardized distribution  $g_\lambda$ :

$$\begin{aligned} \partial_\lambda g_\lambda(x) - \left( \frac{\ddot{\mu}_\lambda}{2\dot{\mu}_\lambda} x + \dot{\mu}_\lambda^{1/2} \right) \partial_x g_\lambda(x) \\ = \left( \frac{\ddot{\mu}_\lambda}{2\dot{\mu}_\lambda} + \dot{\mu}_\lambda^{1/2} x \right) g_\lambda(x). \end{aligned} \quad (16)$$

Here dot means  $d/d\lambda$ . Thus, the shape of the relative entropy minimizer satisfies a (time-dependent, inhomogeneous) *transport* equation. It can be checked that (16) preserves the normalization, mean and variance of  $g_\lambda$  as it should.

The existence of a unique attractor for such a first-order transport equation is somewhat counter-intuitive: we are used to thinking of transport as a non-dissipative process (initial distributions are “moved around” without information being destroyed or created). In contrast with this intuition, we have seen that a large domain of initial conditions  $g_0$  converge to the standard Gaussian  $\phi$  under the transport flow (16). The reason for this behaviour is of course the presence of the “self-referential” function  $\mu_\lambda$  in this equation:  $\mu_\lambda$  is determined by the initial condition  $q = p_0$ , thereby rendering the problem non-linear. In other words, the function  $\mu_\lambda$  captures the shape of the initial distribution in such a way that the time-dependent terms in (16) “erase” this information over time.

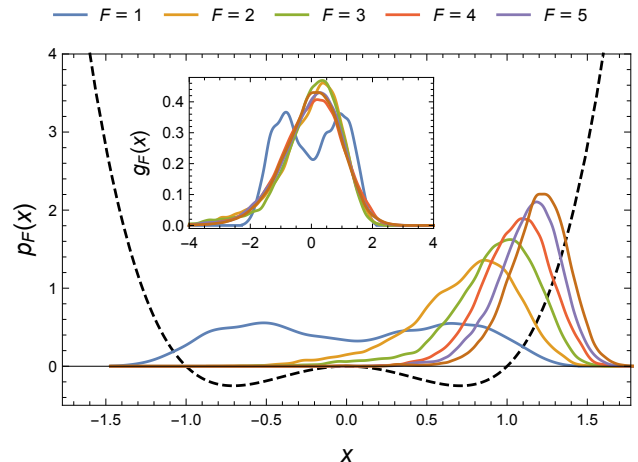


FIG. 2. Simulated equilibrium distributions for a Brownian particle in a Mexican hat potential  $V_0(x) = -x^2 + x^4$  (dashed line) under different applied forces  $F$ . The inset shows the corresponding standardized distributions  $g_F(x)$ , which approach the standard Gaussian as  $F$  increases. Here  $\gamma = 1$ ,  $T = 1$  and  $t \in [0, 100]$  with steps  $\delta t = 10^{-3}$ .

## APPLICATIONS

### Driven Brownian particle

As a straightforward application of our limit theorem, consider the overdamped motion of a Brownian particle in one spatial dimension, *viz.*

$$\frac{dx_t}{dt} = -\gamma V'(x_t) + \xi_t, \quad (17)$$

with  $x_t$  the position of the particle at time  $t$ ,  $V(x)$  a potential,  $\gamma$  a friction coefficient, and  $\xi_t$  is a Gaussian white noise with  $\langle \xi_t \xi_s \rangle = 2\gamma T \delta(t - s)$ . Assume that  $V(x)$  consists of a smooth confining part  $V_0(x)$  and of a constant applied force  $F$ , *i.e.*  $V(x) = -Fx + V_0(x)$ . Then the equilibrium distribution is

$$p_F(x) \propto \exp\left(\frac{Fx - V_0(x)}{T}\right), \quad (18)$$

and the results in the previous sections imply that  $p_F(x)$  must be Gaussian in the limit of large forces  $F$ , irrespective of the background potential  $V_0$ . We illustrate this finding with a Mexican hat potential in Fig. 2.

### Natural selection

Let us now consider a different application in the context of evolutionary dynamics [6]. Darwin’s principle of the “survival of the fittest” may be stated as follows: in a population of replicators such that (i) each replicator

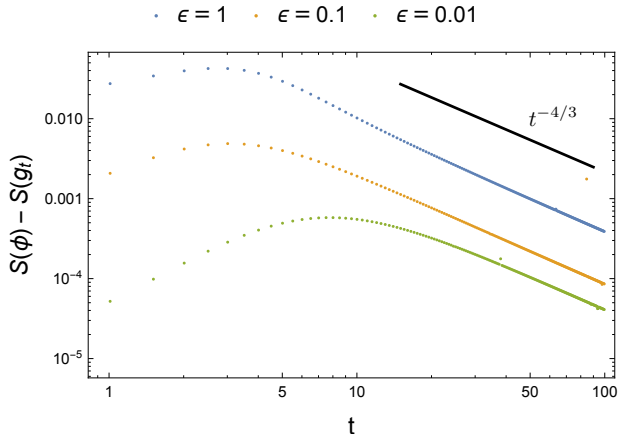


FIG. 3. Entropy is eventually increasing but it is not a Lyapunov function for the transport equation (16). Here the initial conditions are  $p_0(x) \propto \exp(-x^2/2 - \epsilon x^4)$  for three different values of  $\epsilon$ . Note that, oddly, the closer the initial distribution from the Gaussian (*i.e.* the smaller  $\epsilon$ ), the later the standardized entropy  $S(g_t)$  starts increasing towards its limit  $S(\phi)$ .

has a well-defined growth rate (aka “fitness”)  $x$  (*exponential growth*), (*ii*) not every replicator has the same fitness (*variation*), and (*iii*) the fitness of descendants is approximately equal to the fitness of parent replicators (*heredity*), then the descendants of the replicators with maximal fitness will eventually take over the entire population, *i.e.* their relative fraction will converge to one. While originally formulated to account for the evolution of biological species,<sup>4</sup> this principle is applicable in variety of contexts, from molecules to languages to algorithms to firms. The general relevance of natural selection as an evolutionary force is referred to as “Universal Darwinism” [14].

A refinement of the principle of the survival of the fittest is Fisher’s “fundamental theorem of natural selection” [15]. This celebrated result is the observation that (*i – iii*) imply that the mean fitness  $\mu_t$  in the population grows in time as

$$\frac{d\mu_t}{dt} = \sigma_t^2, \quad (19)$$

with  $\sigma_t^2$  the fitness variance at time  $t$ . In particular  $\mu_t$  can never decrease under natural selection. Fisher compared this fact with the second law of thermodynamics,<sup>5</sup>

an analogy which has been hotly debated ever since [16]. Our result above suggests an alternative heuristic connection between evolutionary dynamics and the second law. Instead of its mean and variance, this new connection involves the entropy of the fitness distribution.

Consider indeed a population of replicators such that the density of individuals with growth rate  $x$  is  $p_0(x)$ . Then as a consequence of Darwin’s principles (*i – iii*), we must have after a time  $t$

$$p_t(x) \propto e^{xt} p_0(x), \quad (20)$$

*i.e.* the evolved fitness distribution  $p_t$  is the minimizer of  $D(p_t \| p_0)$  with mean  $\mu_t$  [17]. Thus knowing the initial fitness distribution and the mean fitness at all times is equivalent to knowing the entire fitness distribution at all times. Equivalently,  $p_t(x)$  is the solution of (15) with  $\lambda$  as time  $t$ .

Now, according to the theorem derived above, provided the population is sufficiently large and diverse so that the support of  $p_0$  is effectively unbounded (*i.e.* in a regime of “positive” natural selection [6]), the fitness distribution will by force become Gaussian over time. Moreover a single “conserved quantity” (the  $\alpha$  tail exponent) completely controls the late-time behavior of the evolving population. Such universality implies that natural selection is a *predictive* hypothesis. That such a system-independent prediction are even possible is sometimes disputed by biologists, who tend to emphasize the “contingency” of evolutionary changes rather than its universal statistical structure.

To highlight the similarity between the present limit theorem and the  $H$  and central limit theorems, it is useful to reformulate our main result in terms of entropy. (We recall that both the central limit theorem and the  $H$  theorem are statements about the monotonicity of entropy under the relevant flow—though in the former case this was proved only recently [18]). Under the same assumptions as above, we can show that

$$S(\phi) - S(g_t) \underset{t \rightarrow \infty}{\sim} \frac{(\alpha C)^{1/(\alpha-1)} (2 - \alpha)^2}{\alpha - 1} t^{-\alpha/(\alpha-1)}. \quad (21)$$

We note that this result is superficially similar to Iwasa’s evolutionary  $H$  theorem [19], which identifies a “free fitness function that always decreases in evolution”. However important differences should be emphasized. First, Iwasa’s theorem applies to Markovian models of evolution, and as such it is a result in linear partial differential equations; Eq. (15), by contrast, is a non-linear

<sup>4</sup> Somewhat paradoxically, biological evolution may be the field where natural selection is least strongly established as a dynamical principle. Even condition (*i*) is hard to verify in real populations [11, 12], and it takes experimental engineering to realize exponential replicators in the lab [13].

<sup>5</sup> From [15]: “Professor Eddington has recently remarked that

‘The law that entropy always increases—the second law of thermodynamics—holds, I think, the supreme position among the laws of nature’. It is not a little instructive that so similar a law should hold the supreme position among the biological sciences.”

integro-differential equation without a Markovian interpretation. Second, Iwasa’s theorem involves the relative entropy of the probability distribution with respect to a system-dependent final state. Here, on the other hand, the late-time distribution is universal, resulting in a general statistical prediction of Darwin’s theory of evolution through natural selection. Third, our result applies to the standardized fitness distribution  $g_t$ , not to the fitness distribution  $p_t$  itself. This is more similar to the entropic central limit theorem [18], which is statement about *rescaled* sums of i.i.d. variables, than to Iwasa’s theorem. Fourth, unlike relative entropy for Markov processes, the entropy of  $g_t$  is *not* a Lyapunov functional for the flow (16), see Fig. 3

## CONCLUSION

Minimum relative entropy distributions with a large mean are asymptotically Gaussian when  $\mu \rightarrow \infty$ . We gave a proof of this result in terms of cumulants, but an alternative, direct-space formulation involving a “self-referential” transport equation exists. It would be interesting to understand the dissipative nature of this flow more precisely, for instance by exhibiting a Lyapunov function.

I thank Cédric Villani for a stimulating discussion and for drawing my attention to Ref. [18]. Research at the Perimeter Institute is supported in part by the Government of Canada through Industry Canada and by the Province of Ontario through the Ministry of Research and Innovation.

- 
- \* msmerlak@perimeterinstitute.ca
- [1] S. Kullback, *Information Theory and Statistics* (Wiley, 1959).
  - [2] E. T. Jaynes, Phys. Rev. **106**, 620 (1957).
  - [3] J. Shore and R. Johnson, IEEE Trans. Inform. Theory **26**, 26 (1980).
  - [4] B. Buck and V. A. Macaulay, *Maximum entropy in action*, a collection of expository essays (Oxford University Press, USA, 1991).
  - [5] L. Itti and P. Baldi, Vision research **49**, 1295 (2009).
  - [6] M. Smerlak and A. Youssef, arXiv (2015), 1511.00296.
  - [7] N. H. Bingham, C. M. Goldie, and J. L. Teugels, *Regular Variation* (Cambridge University Press, Cambridge, 1989).
  - [8] Y. Kasahara, J. Math. Kyoto Univ. **18**, 209 (1978).
  - [9] W. Feller, *An Introduction to Probability Theory and its Applications* (Wiley, 1979).
  - [10] L. de Haan and A. Ferreira, *Extreme Value Theory*, An Introduction (Springer Science & Business Media, New York, NY, 2007).
  - [11] G. von Kiedrowski, in *Bioorganic Chemistry Frontiers* (Springer Berlin Heidelberg, 1993) pp. 113–146.
  - [12] I. A. Hatton, K. S. McCann, J. M. Fryxell, T. J. Davies, M. Smerlak, A. R. E. Sinclair, and M. Loreau, Science **349**, aac6284 (2015).
  - [13] M. Colomb-Delsuc, E. Mattia, J. W. Sadownik, and S. Otto, Nat. Communications **6** (2015).
  - [14] R. Dawkins, in *Evolution from molecules to man*, edited by D. S. Bendall (In Evolution from Molecules to Men (1983), pp. 403–425, 1983) pp. 403–425.
  - [15] R. A. Fisher, *The Genetical Theory of Natural Selection*, A Complete Variorum Edition (Oxford University Press, 1930).
  - [16] S. A. Frank, Evolution **51**, 1712 (1997).
  - [17] G. P. Karev, Bull. Math. Biol. **72**, 1124 (2010).
  - [18] S. Artstein, K. Ball, F. Barthe, and A. Naor, J. Amer. Math. Soc. **17**, 975 (2004).
  - [19] Y. Iwasa, J. Theor. Biol. **135**, 265 (1988).